Tools for the Internet

Channeling the Information Tidal Wave

InTEXT

Contents

Executive Summary 1

"Wired" World Requirements 2

Architectural Overview 3

InTEXT Technology Makes All the Difference 3

InTEXT Unique Benefits to Web Site Developers 4

InTEXT SDKs 5

About *InTEXT* Systems 7

Executive Summary

"The Internet and World Wide Web explosion has created a tremendous need for tools that support on-line (live) information, then profile and summarize it based on its content. InTEXT has intelligent Internet tools that help companies best utilize their on-line information assets now and in the future."

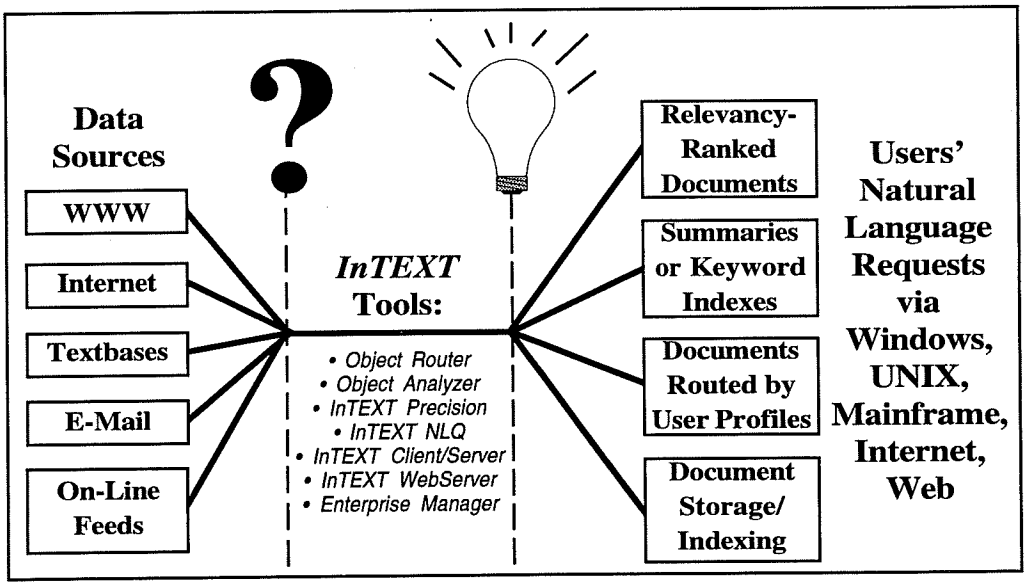
—Karan Eriksson, CEO, InTEXT Systems

The Internet and World Wide Web provide on-line text to users world-wide. Organizations creating Web sites face three crucial requirements for leveraging their Web site investment: produce a scalable, easy-to-maintain Web site for cost efficiency; parse on-line (live) data, without indexing, to have timely information; and use tools with content understanding to reduce information overload.

The best way to ***produce a scalable, easy-to-maintain Web site*** is through an open architecture. An open textbase development architecture, such as *InTEXT*'s Heuristic/Learning™ architecture, supports extensive scalability such as incremental user increases, add-on technology, macro functionality, APIs, and extensibility to document management, relational database, and application development environments.

Organizations must also be able to ***profile live, on-line information*** so that it remains timely. True on-line profiling includes the ability to parse, summarize, and route information without needing to index it. *InTEXT*'s technology automatically summarizes documents to reduce network overhead and routes live newsfeeds to users based on their profiles.

Technology using ***content understanding*** is key to accessing the *right* information quickly and precisely. *InTEXT*'s intelligent tools skim-read the surface structure of documents and assess their information content, or, "aboutness". These tools also self-tune to document collections, keeping organizations abreast of new and evolving information.



InTEXT Systems' technology sifts through on-line data from multiple sources and parses relevant information to users in real-time (no indexing required) or into text storage

"Wired" World Requirements

The World Wide Web is gaining ground at a colossal rate, is generating huge interest, and looks to be unstoppable. To have a successful Web site, organizations should be aware of the following requirements:

The need for IT/developers to meet the organization's expectations—Since organizations expect their IT department or systems integrators to make them a competitive Web site quickly and inexpensively, IT/developers have several requirements: 1) security, 2) cheap start-up costs, 3) fast application development, 4) distributed architecture, 5) scalability, 6) minimal tuning, 7) minimal storage overhead, 8) easy maintenance, 9) reliability, and 10) proven, yet innovative, technology.

Demand for a scalable architecture and a comprehensive set of interoperable tools—In creating Web sites, organizations begin with a small number of users and expand incrementally. At the same time, they start with few machines and simple configurations that change with time. Organizations need tools that are built on an open architecture—one that supports industry standards, user and configuration changes, new platforms, new software, and many other unforeseen changes. A good architecture helps to future-proof Web site investments.

The need for organizations to utilize intellectual assets—The proliferation of productivity tools at the desktop (e.g., windows, spreadsheet, word processor, database) distributes organizations' disparate intellectual assets across PCs and workstations in on-line document form. It is crucial for organizations to access this investment quickly, easily and seamlessly.

The demand for tools to make on-line text more accessible, usable, and understandable—The influx of electronic information through news groups, e-mail and other on-line sources leaves users burdened with vast amounts of data and no time to read it. It is crucial to have technology that can skim-read text, comprehend what it is talking about in real-time, and automatically summarize and/or route the information for users, all based on content.

The need to have access to both historical and new information—Organizations' historical information is a valuable asset that should be easily accessible. At the same time, users need to capture the new information flowing through news groups, e-mail, Usenet, etc. Text management and retrieval technology must provide access to both information stored on local databases and information flowing daily across user desktops.

Architectural Overview

Document management and retrieval has evolved for over two decades, and is evolving again to meet the needs of organizations confronting Internet and Web publishing. With the new requirements created by the Internet and Web, organizations need to support two key data sources: stored and live information.

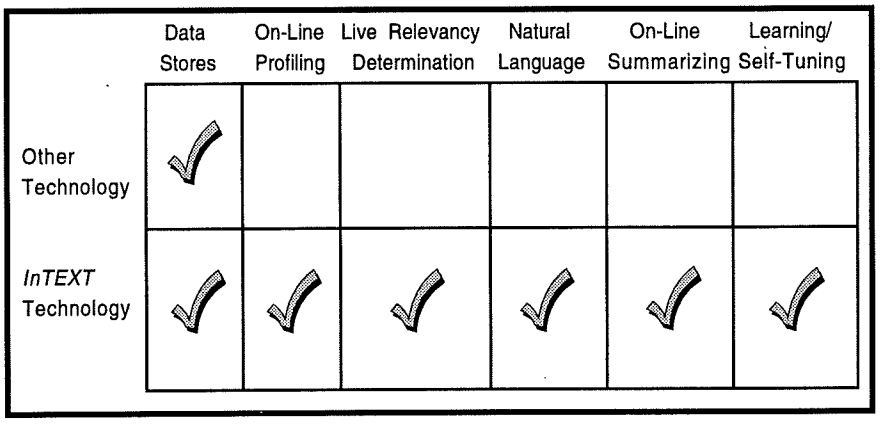
Stored Information: This consists of large bodies of textual information stored in many different forms (e.g., WAIS, full-text retrieval index, RDBMS BLObs, World Wide Web). While this information is no longer timely, it is considered relevant when a user needs it, thus is considered an asset when it can be located and used. *InTEXT*'s Heuristic/Learning architecture provides fast and accurate access, retrieval, and storage of organizations' information assets.

Live Information: This consists of live information feeds (e.g., Usenet News, Email, news wires). Crucial information is most often the timeliest. Users need daily access to crucial data without receiving a tidal wave of unimportant information. Because it can read the surface structure of text and comprehend what it is talking about in real-time, the *InTEXT* Heuristic/Learning architecture can discern whether to route incoming information to specific users immediately or to store it for future retrieval.

InTEXT's Heuristic/Learning technology supports both stored and live information, creating a powerful and flexible Internet and Web document management, routing, and retrieval solution.

InTEXT Technology Makes All the Difference

InTEXT's Heuristic/Learning technology can skim-read on-line documents, determine their relevancy, create summaries, and self-tune to new information collections in real-time, all in response to users' natural language profiles or queries.



InTEXT Systems' solutions are based on a powerful content analysis, or, Heuristic/Learning architecture. A Heuristic/Learning architecture uses skim-reading, comprehension, and self-tuning techniques to understand the content of information. It determines incoming information's relevance to other documents; discovers its key words, sentences and phrases; and dynamically routes documents to users based on content relevancy.

In general, *InTEXT*'s *heuristics* are based on techniques that authors use to make their point in their writing, such as repetition, structuring into paragraphs, use of titles, etc. Unlike architectures that use only statistical

The Heuristic/Learning Architecture compares users' interest profiles to the content of on-line data and routes the most relevant information to users as complete or summarized documents.

methods such as frequency analysis, the heuristics employed by *InTEXT* "treat text as text". So heuristics retain the author's original meaning rather than altering a document's theme.

The *learning* component within the architecture tunes the heuristics to the stream of text being processed by the software. Often, document management architectures require front-loading, such as creating static structures and hand-crafting weights before being able to use the product. With *InTEXT*'s technology, the heuristics are built in, and the tools are ready to use immediately. Further, the heuristics self-tune to streams of information, thus learning from text and document content.

InTEXT Unique Benefits to Web Site Developers

Through the Heuristic/Learning architecture, *InTEXT*'s technology provides several unique benefits to organizations creating Web sites:

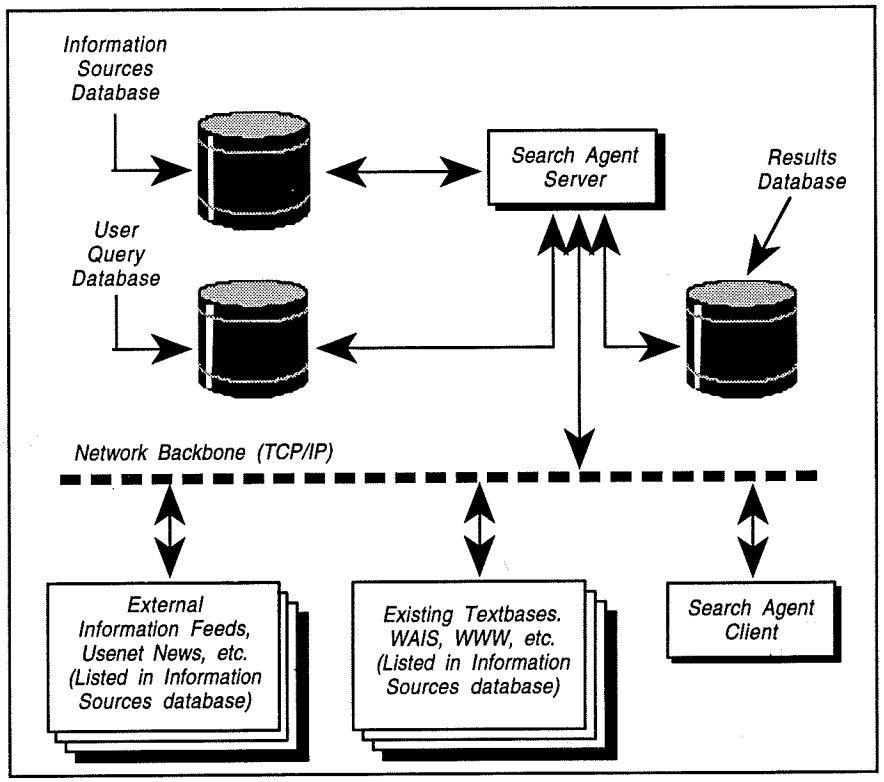
Feature	Purpose	Benefit
Real-time handling of on-line text	Automatically pass information to users or to storage—*no indexing required*	• Crucial information remains timely • Have competitive information immediately
Content understanding	Instantly know what on-line information is "about"	• Assess information relevance on the fly • Access and use the *right* information immediately
Information gathering & distribution	Automatically find all relevant information, based on its content (e.g., live text streams, stored documents)	• Reduce information overload • Self-tune to information streams for timely access to current information
Document summarization	Automatically summarize documents, (e.g., keywords and phrases, abstracts, titles)	• Reduce information overload • Reduce network traffic • Improve responsiveness
Hyperlink generation	Automatically create Web hyperlinks (i.e., the major navigation form of the Web)	• Cost effectiveness • Stylistic consistency • Automatic hyperlink creation/updating
Free-form querying	Accept user queries in standard, free-form English for content-based searches	• Users get the on-line information they need *without* having to learn complicated query structures

InTEXT Software Development Kits

These powerful capabilities are delivered as separate tools or as software development kits (SDKs):

- *User profiling*
- *On-line information monitoring and profiling*
- *Content understanding*
- *Unlimited support for agent filters and simultaneous users*

InTEXT Object Router SDK—Sorts on-line documents with respect to their content—no indexing required. The Router allows users to subscribe to profiles that define their information needs. Each profile in the system, and there may be many thousands, acts as an agent which searches for relevant content. This content retrieving is achieved through a set of terms that are matched against the words and phrases of documents processed by the Router. The Router performs all document processing in real-time.



The Object Router intelligent content agent toolkit routes on-line and stored information to users in real-time—no indexing required. This is just one of many possible Object Router applications.

- *Real-time summaries*
- *High performance analysis & content understanding*
- *Automatic hyperlink and keyword generation*
- *Automatic Web page generation (HTML)*

InTEXT Object Analyzer SDK—Provides intelligent summary information about documents automatically. The purpose of the Object Analyzer is to identify the major information content of a document, either for the purposes of indexing it, or to assist a reader in comprehending its contents more rapidly. The important information is presented as a relevant summary. The Object Analyzer retains the original context of the document to maintain the author's meaning while creating 1 to 99 percent summaries of documents. The Object Analyzer can automatically generate hypertext links for creating webbed documents.

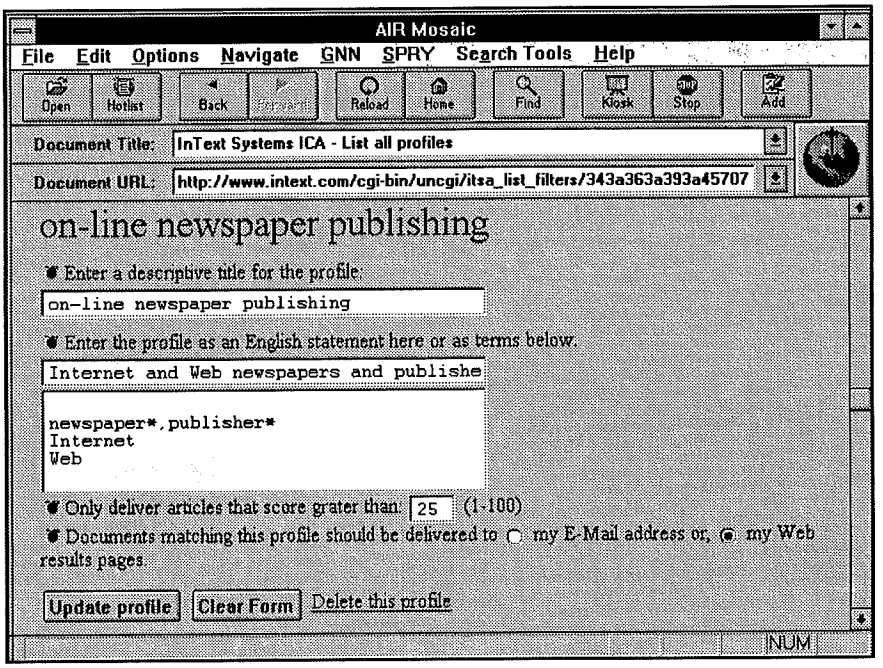
- *Significantly reduces index sizes*
- *Converts native word processing documents to HTML*

InTEXT Precision SDK—Utilizes a PreciseScoping technology that generates a significantly smaller full-text index than any other commercial product. Precision automatically determines documents' most content-bearing, relevancy-weighted words and phrases and creates summarized documents. These summarized documents enable a significant increase in precision while producing indexes that are 5 to 10 times smaller. Precision also creates logical structure i.d.'s and SGML, HTML, and keyword tags.

- *Simple application installation—no dictionaries required*
- *Domain independent*
- *Supports natural English & fuzzy queries*

InTEXT NLQ SDK—Takes short passages of text (typically one or two sentences) and generates a structured form suitable to pass either to a document retrieval system as a query, or to a document filtering system as a definition of a user's interest. NLQ removes the need for users to learn any query syntax, often the biggest obstacle to end-user usage of document retrieval. Further, NLQ allows the application to support the use of a "seed sentence" to create a query from a piece of text copies out of a document found in the collection.

The InTEXT NLQ is a powerful toolkit for submitting free-form queries across organizational, Internet, and Web textbase servers.



- *Content-aware, distributed document management, storage & retrieval*
- *Desktop/LAN/ Mainframe connectivity*

InTEXT Enterprise Manager SDK—Provides the only "desktop-to-LAN-to-mainframe" distributed document storage, management, and retrieval solution available today. With Enterprise Manager, users have full access to Document Management functions such as full check-in/ check-out management and document migration across LANs and WANs—all from their desktops.

- *Content-based search*
- *Relevancy-ranked retrieval*
- *Native document display and output*
- *Plain English and boolean searches*
- *Enterprise-wide scalability*
- *WAIS and Z39.50 standardized*

InTEXT WebServer SDK—Provides advanced full-text document storage and retrieval for organizational, Internet, and Web servers. *InTEXT* WebServer provides content-based retrieval of documents, lists documents in relevancy-ranked order, and supports both Boolean and Free-Form English queries. Its indexes contain the precise position of each word, allowing phrase and proximity searches to be made without having to scan the original document—a key requirement for long documents.

The *InTEXT* WebServer allows information providers and corporate publishers a straightforward way to provide Internet and Web access to their documents. Complying with WAIS and Z39.50 query protocols, WebServer supports Web browsers, such as Netscape and Mosaic, and WAIS-compliant Internet clients. Its databases are fully compatible with third party textbases such as STATUS™.

About *InTEXT* Systems

InTEXT Systems delivers advanced software products and technologies for content-based routing, retrieval, development, and presentation for mission-critical, workgroup, Internet, and World Wide Web applications. A company of CP Software Group, *InTEXT* is backed by over 12 years of focused research and development in the areas of intelligent analysis, routing, and retrieval.

Headquartered in San Francisco, Calif., *InTEXT* offers worldwide sales, technical support, and consultation services. *InTEXT* maintains regional offices throughout the United States, Australia, Asia, and the United Kingdom to support its sales staff, value-added marketers, and software product distributors. *InTEXT* continues advancing its software product suite through its United States and Australian-based R&D laboratories.

InTEXT's routing, analysis, and retrieval technology is used by leading companies such as American Express, Island Software, Uniplex Software, Asymetrix, Electric Power Research Institute, State of California, State of Tennessee, Commonwealth Edison, EXXON, The Wollongong Group, McDonnell Douglas, National Semiconductor, Cybergrahic Systems, Pacific Bell, UPJOHN, the Australian Department of Defense, and many more.

InTEXT's toolkits and intelligent agent technologies are available in several environments, including VAX/VMS, MVS/CICS, VM/CMS, MS/DOS, OS/2, Windows, SunOS, Solaris, AIX, and HP/UX. Call today for pricing and information. Phone numbers are listed on the back of this white paper.

InTEXT
Intelligent Internet Tools

CP Software Group–World Headquarters
715 Sutter St
Folsom CA 95630
Telephone: +1-916-985-4445
Facsimile: +1-916-985-3557

Europe
Telephone: +44-1793 542099
Facsimile: +44-1793 619243

InTEXT Systems–World Headquarters
120 Montgomery St
Suite 450
San Francisco, CA 94104
Telephone: +1-415-391-5290
Facsimile: +1-415-391-4996

Australia/New Zealand
Telephone: +616-283-6877
Facsimile: +616-285-4316

M7029-C 07/95